

This document is the publisher's version of:

Nida, N., Yousaf, M. H., Irtaza, A. y Velastin, S.A. (2019). Instructor activity recognition through deep spatiotemporal features and feedforward Extreme Learning Machines. *Mathematical Problems in Engineering*, 2474865.

DOI: <https://doi.org/10.1155/2019/2474865>

© 2019 Nudrat Nida et al.



This work is licensed under a [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Research Article

# Instructor Activity Recognition through Deep Spatiotemporal Features and Feedforward Extreme Learning Machines

Nudrat Nida,<sup>1</sup> Muhammad Haroon Yousaf <sup>1</sup>, Aun Irtaza,<sup>2</sup> and Sergio A. Velastin<sup>3,4,5</sup>

<sup>1</sup>Department of Computer Engineering, University of Engineering and Technology, Taxila, Pakistan

<sup>2</sup>Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan

<sup>3</sup>Cortexica Vision Systems Ltd., UK

<sup>4</sup>Queen Mary University London, UK

<sup>5</sup>University of Carlos III Madrid, Spain

Correspondence should be addressed to Muhammad Haroon Yousaf; [haroon.yousaf@uettaxila.edu.pk](mailto:haroon.yousaf@uettaxila.edu.pk)

Received 30 November 2018; Revised 8 March 2019; Accepted 25 March 2019; Published 30 April 2019

Academic Editor: George A. Papakostas

Copyright © 2019 Nudrat Nida et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Human action recognition has the potential to predict the activities of an instructor within the lecture room. Evaluation of lecture delivery can help teachers analyze shortcomings and plan lectures more effectively. However, manual or peer evaluation is time-consuming, tedious and sometimes it is difficult to remember all the details of the lecture. Therefore, automation of lecture delivery evaluation significantly improves teaching style. In this paper, we propose a feedforward learning model for instructor's activity recognition in the lecture room. The proposed scheme represents a video sequence in the form of a single frame to capture the motion profile of the instructor by observing the spatiotemporal relation within the video frames. First, we segment the instructor silhouettes from input videos using graph-cut segmentation and generate a motion profile. These motion profiles are centered by obtaining the largest connected components and normalized. Then, these motion profiles are represented in the form of feature maps by a deep convolutional neural network. Then, an extreme learning machine (ELM) classifier is trained over the obtained feature representations to recognize eight different activities of the instructor within the classroom. For the evaluation of the proposed method, we created an instructor activity video (IAVID-1) dataset and compared our method against different state-of-the-art activity recognition methods. Furthermore, two standard datasets, MuHAVI and IXMAS, were also considered for the evaluation of the proposed scheme.

## 1. Introduction

A tremendous amount of video sequences is generated every day from CCTV cameras, YouTube, surveillance systems, the entertainment industry, and academic institutes. Manual analysis of visual information is time-consuming and error-prone. In an era of advanced computer vision technology, it is possible to use automatic visual understanding methods to understand the visual semantics of the classroom. Teaching effectiveness is a fundamental concept in contemporary education, valued by academic institutions as a goal on their own right. Some researchers have explored human pose recognition techniques using handcrafted features for estimating the instructor's activities in the classroom [1–4] walking, writing, pointing towards the board, standing,

and addressing and pointing towards presentations, respectively. Silhouette representation is often computationally less expensive but demands precise segmentation of human silhouettes for pose estimation and such techniques have primarily focused on handcrafted representations of spatial information. The temporal anchoring of spatial frames is not incorporated in these methods.

The literature on Human Activity Recognition (HAR) can be grouped into traditional (handcrafted) and deep learning action representation. Handcrafted spatiotemporal features encode the appearance and movement profile of actor for better action prediction [5–7]. For instance, Dollar et al. [5] propose the mapping of 2D to 3D spatiotemporal interest points as cuboid descriptions for actions prediction, while Wang and Schmid [7] establish dense motion trajectories

(iDT) by computing the camera movement information. There are various types of spatiotemporal features to generalize action recognition including, spatiotemporal features [1], 3D-SIFT [8], HOG3D [9], extended SURF [10], iDT [7], histogram of optical flow (HOF) [11], and motion boundary histogram (MBH) [12]. Describing the iDT with MBH, HOG, HOF have shown the better prediction of activities on benchmark datasets (UCF101 [13], HMDB [14] and THUMOS [15]). However, sometimes the trajectory information is degraded due to large variations among action categories and it can be argued that they are not suitable for realistic HAR tasks [16].

Recently, deep learning approaches for HAR modified the 2D CNN models to capture 3D spatiotemporal action representation. Ji et al. [17] modify the 2D-CNN into 3D-CNN using neighboring frame information. However, the performance of 3D-CNN is comparable with 2D-CNN, where 2D-CNN was modeled for image category recognition tasks [18]. Simonyan et al. [19] suggest a two-stream ConvNet to encode frame appearance information combined with motion information. The two streams ConvNet performance is comparable with the performance of iDT on UCF101 and HMDB51. The combination of handcrafted and deep features have improved action prediction rate [18], due to the fact that sparse spatiotemporal handcrafted features are encoded into deeper representation and accurately recognize the activities. The 2D and 3D-CNN for activity recognition are trained by back propagation methods to reduce the classification loss and sometimes suffer from overfitting. However, for action recognition applications, the amount of training data is small to establish a generic model. Some of the standard HAR datasets are UFC-101 [13], comprising 13K videos of 101 action classes, MuHAVI-Uncut [20], which consists of 2898 videos of 17 classes and HMDB-51 [14] dataset consisting of 3.7 K videos of 51 categories.

The visual information of the classroom holds significant metaphors to provide genuine feedback for lecture effectiveness [21, 22]. We believe that computer vision techniques are beneficial to automate a fair instructor's [3] learning model to recognize eight activities of an instructor within the lecture room. In the proposed technique, instructor silhouettes are segmented from the static background using graph-cut segmentation. Silhouettes of each video frame are used to encode spatial and temporal dynamics of instructor activities through motion profiles. These motion profiles store the spatiotemporal instructor's contextual information from each video sequence in single templates. Then, these motion profiles are used to compute deep features by applying deep convolutional operations and induce nonlinearity among the deep spatiotemporal representation. Then, these learned features are presented to Extreme Learning Machines (ELM) to generate a feedforward model for instructor activity recognition. An overview of the proposed model is shown in Figure 1.

At times, transfer learning of pretrained CNN models suffers from poor performance and overfit due to lack of data. However, our proposed technique learns deep spatiotemporal action representation and performs fast and accurate action recognition, even with limited data. We have elaborated this contribution using our IAVID-I dataset. The proposed technique works on feedforward learning and performs better

than backpropagation CNN models for action recognition. Moreover, the motion profile effectively reduces the video's spatiotemporal and computational complexity.

The deep features capture the high-level discriminative representation from the motion profile for prediction of instructor actions. The performance of the proposed technique has been evaluated on our recorded single-view IAVID-I dataset for instructor activity recognition and also on benchmark multiview activity recognition datasets (MuHAVi, IXMAS). However, as far as we know, such a feedforward model has not been reported before for action prediction task. The proposed technique achieved higher prediction scores on MuHAVI-Uncut with 2989 videos compared to state-of-the-art techniques.

Our contributions are summarized as follows:

- (i) We propose a new feedforward learning model for fast and accurate instructor activity recognition.
- (ii) The fast feedforward proposed technique can learn deep features from any kind of CNN model.
- (iii) The technique is able to be applied for silhouette-based activity recognition applications.
- (iv) We have shown that the proposed approach can be used for multiview human action recognition. This contribution is explained further in the experiments section using a standard multiview HAR dataset (MuHAVI-Uncut).

## 2. Proposed Method

The proposed technique consists of a three-step process, as shown in Figure 1. First, we extract the instructor silhouettes  $f$  of each video frame to generate cumulative spatiotemporal instructor's motion profile  $M_f(x, y, t)$ . Then, we learn the deep spatiotemporal features  $x$ . Then, we present these spatiotemporal deep features to an ELM for instructor's activities recognition into eight action classes.

**2.1. Silhouettes Segmentation and Spatiotemporal Motion Profile Formation.** The instructor silhouettes  $f$  are segmented from RGB videos using graph-cut segmentation [23]  $p_{ab}$  of video frame generating a corresponding graph vertex  $v_{ab}$  of the graph. The foreground silhouettes  $f$  and static background  $B$  of lecture room are presented as two additional vertices. The weights on the links between the pixel vertices and foreground  $f$ , background  $B$  are derived from the difference between the background and the current frame at the corresponding pixel,  $q_{ab}$ .

$$\omega(f, p_{ab}) = q_{ab} \quad (1)$$

$$\omega(p_{ab}, B) = 2\varphi - p_{ab} \quad (2)$$

where  $\varphi$  is a threshold parameter that determines the association of  $p_{ab}$  with instructor silhouettes  $f$  and lecture room background  $B$ . The instructor silhouettes  $f_{(1,2,\dots,N)}$  segmented from each video frame are used to generate a single instructor

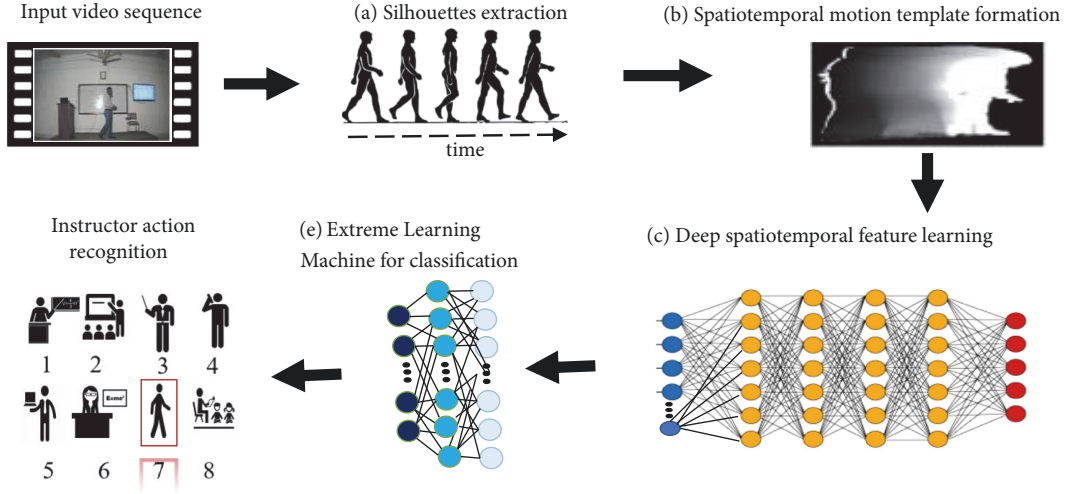


FIGURE 1: Overview of the proposed technique for instructor activity recognition.

motion profile  $M_f$  to encode spatiotemporal movement information at time  $t$ :

$$M_f(a, b, t) = \begin{cases} \tau & \text{if } f(a, b, t) = 1 \\ \max(0, f(a, b, t-1) - 1) & \text{otherwise} \end{cases} \quad (3)$$

Here,  $\tau$  is a total number of frames to generate  $M_f(a, b, t)$  for every action video. All the resulting  $M_f(a, b, t)$  are normalized and rescaled to predefined dimensions for further presenting to deep CNN models.

**2.2. Spatiotemporal Deep Feature Learning.** After obtaining the  $M_f(a, b, t)$ , deep representations  $x$  of instructor actions are generated from  $M_f(a, b, t)$  through CNN. In our algorithm, we have adapted Alexnet [24], and VGG19 [19] are denoted as  $x_{17}$ ,  $x_{20}$ , and  $x_{23}$  (extracted from Alexnet with 1x4096, 1x4096, 1x1000 dimensions),  $x_{39}$ ,  $x_{42}$ , and  $x_{45}$  (extracted from VGG19 with 1x4096, 1x4096, 1x1000 dimensions). The visual representation of deep spatiotemporal features is illustrated in Figure 2. The  $x$  subscript represents the layer depth used for computation of spatiotemporal features.

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (4)$$

$x$  are normalized through the min-max normalization algorithm (eq. (4)). Implementation and observed results are discussed in the Results section.

**2.3. Extreme Learning Machine as a Classifier.** The extreme learning machine is a feedforward learning algorithm using a single layer of the neural network and usually known as Single Layer Feedforward Neural Network (SLFN) [24, 25]. In this work, we investigate this to recognize instructor activities,

something not reported in the literature to date [25, 26] which is used to predict a single output unit to classify instructor activity recognition problem using  $L$  hidden nodes described as

$$y_L(x) = \sum_{j=1}^L \beta_j H_j(x) \quad (5)$$

where  $\beta_i = [\beta_1, \beta_2, \beta_3, \dots, \beta_L]$  are output weights between  $L$  hidden nodes and output vectors,  $Y = [Y_1(x), Y_2(x), \dots, Y_L(x)]$ . The classification decision function of ELM with logistic sigmoid and hyperbola tangent sigmoid activation function is expressed as follows:

$$y_L(x) = \text{sig}(H_i(x) \beta_i) \quad (6a)$$

$$y_L(x) = \tanh(H_i(x) \beta_i) \quad (6b)$$

The logistic sigmoid transforms the input at each hidden neuron and generates a nonlinear output within the 0-1 interval, using the expression (7a). Another activation function is hyperbolic tangent ‘tanh’ function and its output is within [-1, 1], using the expression (7b).

$$\text{sig}(x) = \frac{1}{1 + e^{-x}} \quad (7a)$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (7b)$$

Selection of optimal training parameters is effective in reducing the ELM classification error [25, 26]. The ELM’s input weights are produced randomly using any continuous distribution function. However, the weights at output nodes are produced using a linear system of the minimum norm. In the proposed technique,  $x$  is an  $N \times D$  matrix of deep spatiotemporal features of  $D$  dimension and  $N$  is the number of training samples.  $w$  is a  $D \times L$  matrix and represents the link between the ELM’s input layer and ELM’s hidden layer.

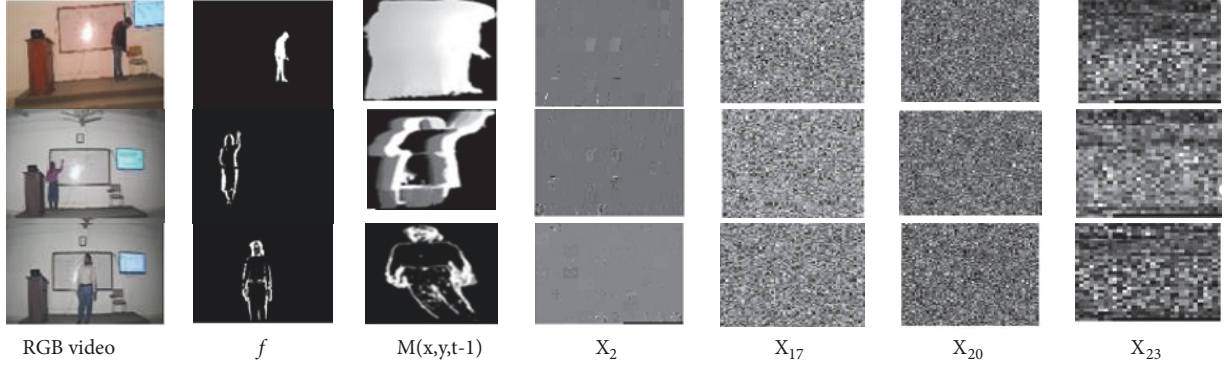


FIGURE 2: Examples of  $x$  feature extracted: (from left to right) original RGB video frames of IAVID-I dataset, instructor silhouettes, motion templates, and  $x$  descriptor, respectively.

$b$  is a bias  $N \times L$  matrix.  $H$  is an  $N \times L$  matrix known as hidden matrix, where  $G(\cdot)$  is a continuous function satisfying ELM universal approximation capability theorem and it is a piecewise nonlinear function. In our proposed technique, we have evaluated  $G(\cdot)$  with a logistic sigmoid and a hyperbola tangent sigmoid activation function. The output weight matrix is  $\beta$  of  $L \times C$  dimensions.

$H^\dagger$  is a Moore-Penrose generalized inverse matrix of  $H$ . The  $T$  matrix is of dimension  $N \times C$  and referred to as target label matrix and holding label vectors in One-Hot encoding scheme for training examples, where  $C$  is the number of instructor activity classes. ELM optimizes the classification process to target generalized performance with minimum training error and norm of output weights using the following objective function:

$$f_o = \underset{\beta_i}{\text{minimize}} \left\{ \|H\beta_i - T\|^2 \right\} \quad (8)$$

Here,  $H$  is the output matrix of the hidden layer as

$$H = \begin{bmatrix} H_1(x_1) & \cdots & H_L(x_1) \\ \vdots & \ddots & \vdots \\ H_1(x_N) & \cdots & H_L(x_N) \end{bmatrix} \quad (9)$$

To minimize the norm of output weights  $\|\beta_i\|$  is achieved through maximizing the margins for strengthening the decision boundary among the eight instructor's action classes within feature representation ( $2/\|\beta_i\|$ ) of ELM, using a minimal least square method as

$$\beta_i = H^\dagger T \quad (10)$$

$H^\dagger$  is the generalized Moore-Penrose inverse matrix calculated through orthogonal projection and the single value decomposition method. The working of ELM in the proposed technique is expressed in Algorithm 1.

### 3. Results and Discussion

In this section, we describe a series of tests performed to evaluate our approach. The following techniques are applied:

- (i) Examine the impact of ELM's hidden nodes for action recognition.
- (ii) Quantitative analysis.
- (iii) Comparison with other state-of-the-art methods.

**3.1. Datasets.** Our investigation includes three different action recognition datasets: our recorded single-view dataset IAVID-I and two standard multiview datasets (MuHAVI-uncut and IXMAS), to evaluate the performance of proposed technique. Some sample action frames are shown in Figure 3. These datasets are described as follows in Figure 3.

**3.1.1. IAVID-I.** We have constructed a dataset of Instructor Activity Video Dataset-I IAVID-I to evaluate the proposed scheme. Twelve actors participated in data recording in realistic lecture room environment focusing on the stage. There are 100, 24-bit RGB videos having 1088x1920 high-resolution. 12 subjects perform the 8 instructor actions and so approximately 12 instances of each action class are present in the dataset. The dataset comprises the following actions: interacting or idle, pointing towards the board, pointing towards the screen, using a mobile phone, using a laptop, reading notes, sitting, walking, and writing on the board, as demonstrated in Figure 3. IAVID-I, publicly available for academic research, is the first attempt with the primary goal to contribute resources in instructor activity recognition. Our dataset will support researchers to test their algorithms for understanding the semantic information within the lecture room. IAVID-I can be a valuable source for algorithm assessment, evaluation, and comparison with state-of-the-art methods.

**3.1.2. MuHAVI-Uncut.** The MuHAVI-uncut dataset is a multiview activity recognition dataset. It contains 17 activities performed by 14 actors at multiple durations. The 8 CCTV cameras were mounted at  $45^\circ$  view difference to capture an action sequence. The MuHAVI-Uncut dataset is a large video dataset (2898 videos) and has segmented single actor's silhouettes.

**3.1.3. INRIA Xmas (IXMAS).** INRIA Xmas Motion Acquisition Sequence (IXMAS) contains 12 activities (cross arms,



**Input:** Deep spatiotemporal features  $x$ , target label  $T$ , number of hidden nodes  $L$ , activation function  $G$ . Let,  $w$  be the weight between ELM input layer and hidden layer,  $b$  is biased vector,  $\beta$  is output weights,  $G$  is the ELM activation function,  $Y$  is the predicted output vector,  $H$  output matrix of hidden layer,  $H^+$  is the generalized Moore-Penrose inverse matrix.

**Output:** parameters of ELM,  $w$ ,  $b$ ,  $\beta$ , and prediction response  $Y$ .

**Generate randomly**  $w$  and  $b$

**Compute**  $H=G(xw + b)$

**Compute**  $\beta=H^+T$ .

**Compute**  $Y=H \beta$

**Return**  $w$ ,  $b$ ,  $\beta$ ,  $Y$

ALGORITHM 1: The ELM algorithm for instructor activity recognition.

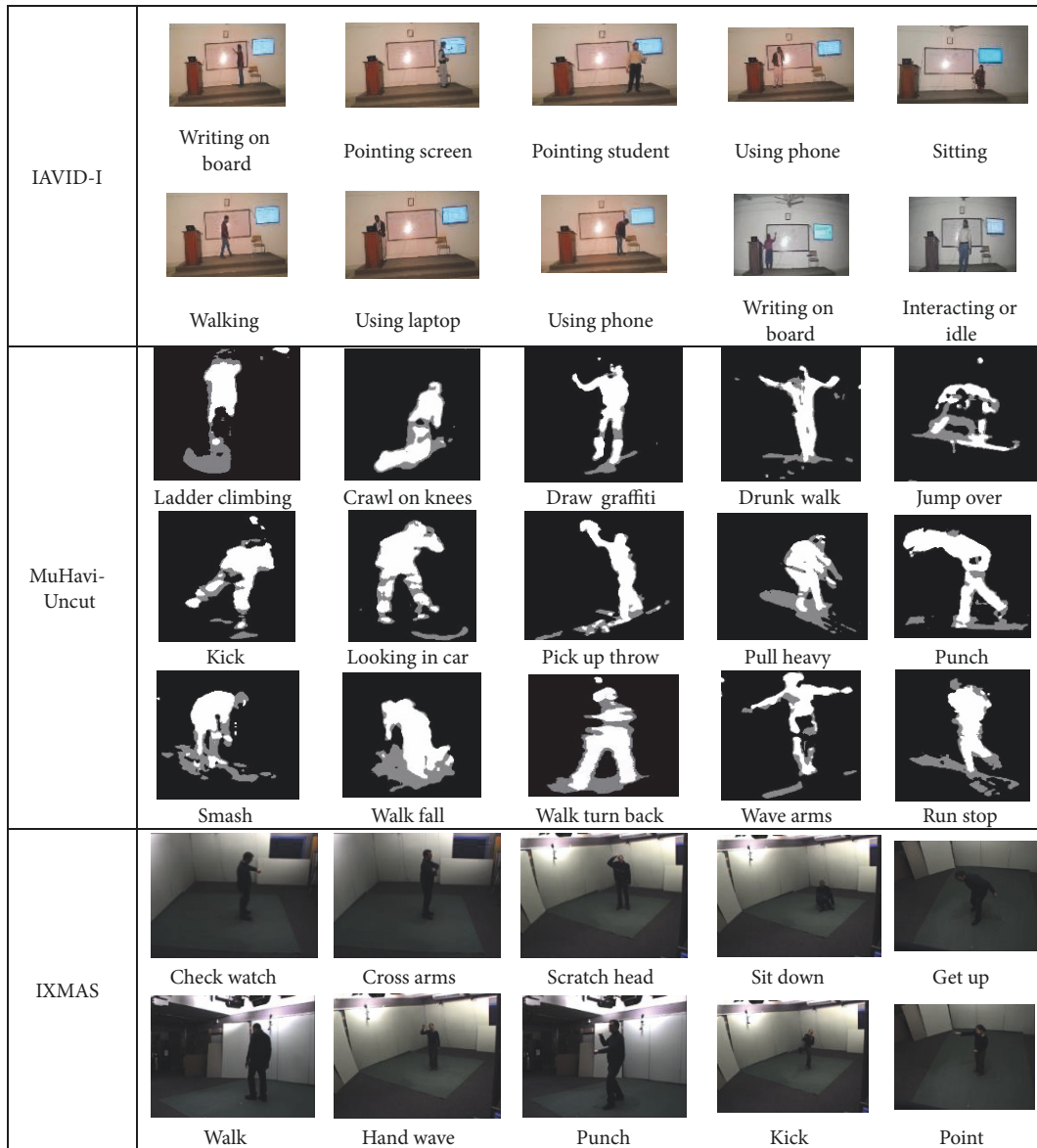


FIGURE 3: Some sample video frames from single-view IAVID-I and multiview MuHAVI-uncut and IXMAS activity recognition datasets.

check watch, sit down, starch head, get up, wave, walk, turn around, kick, punch, pointing, and pick up). Twelve actors perform these activities three times. The dataset was captured from five different views. The frame resolution of each video sequence is 390x291 pixels.

The proposed technique requires actor's silhouettes to form motion templates, as better segmented silhouettes forms better MHIs, and, therefore, MuHAVI-uncut and IXMAS are the most suitable silhouettes datasets for evaluation of proposed technique. Another benefit for proposed technique's evaluation is that MuHAVI-uncut and IXMAS allow us to examine the performance of action prediction for a multiview setting and all the actions in MuHAVI-Uncut and IXMAS dataset are performed by a single actor with a static background, a similar scenario to a single instructor demonstrating in the class.

**3.2. Experimental Setup.** Leave one actor out (LOAO), leave one camera out (LOCO), and leave one sequence out (LOSO) validation schemes are employed in our experiments to evaluate the performance of the proposed model. These schemes define the training and testing splits. For example, in LOAO all the action sequences of one actor are used as a testing and the remaining are used for training. This process is repeated for all the actors and the average performance of the system is recorded. Similarly, in LOCO, action sequences from one camera view are used for testing while the remaining sequences are used for training. In LOSO, all the action sequences are considered for training except for one sequence that is used as a testing sample. The reported accuracies are the average values for each of the experiments.

**3.3. Impact of Number of ELM Hidden Nodes.** ELM is not dependent upon iterative or backpropagation approaches to adjust the weights and bias of SLFNs, rather it analytically estimates the suitable SLFN's parameters, using universal approximation capability with random hidden nodes to establish a generalized model for learning. However, the selection of the number of hidden nodes as a training parameter is effective in reducing classification error [25, 26] behavior of the proposed system.

The parameter 'number of ELM hidden nodes' was chosen using a grid search technique after empirical analysis of the proposed technique on IAVID-I, MuHAVI-uncut and IXMAS single and multiview dataset within the interval of [100-2000]. We have empirically examined the deep spatiotemporal features from the various depths of two types of CNN models (i.e., Alexnet and VGG19) and, in light of our observations,  $x_{17}$  and  $x_{39}$  performed better action recognition. It can be noticed from Figure 4 that deep representation from  $x_{17}$  performed better recognition for both kinds of decision functions. However,  $x_{39}$  slightly shows variation in performance. The sigmoid decision function performed better, as compared to hyperbola tangent sigmoid for a given deep spatiotemporal representation  $x$ . The best number of ELM hidden nodes for IAVID-I dataset is 1300 nodes for  $x_{17}$  and 500 for  $x_{39}$ . However, for MuHAVI-uncut and IXMAS dataset the optimal number of ELM hidden nodes is chosen as 129,060 for  $x_{17}$  and  $x_{39}$  features.

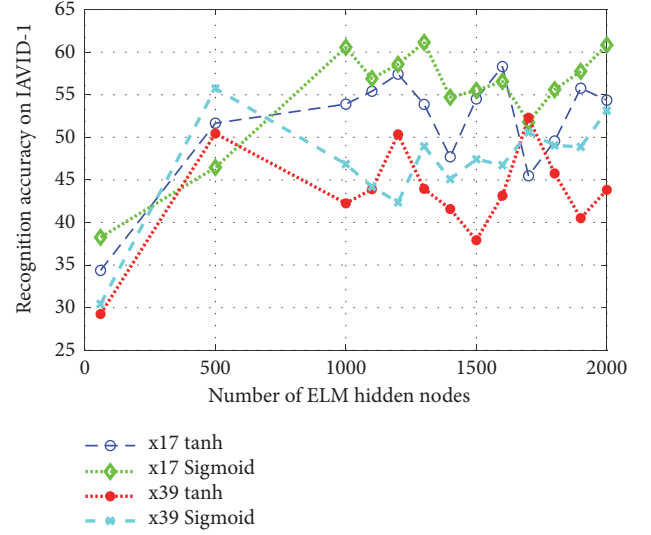


FIGURE 4: Impact of the number of ELM's hidden nodes on activity recognition using logistic sigmoid and hyperbolic tangent sigmoid decision functions on the IAVID-I dataset.

Some of the key findings observed from this experiment are listed as follows:

- (i) The choice of ELM activation function has a significant role in improving the recognition rate. In this experiment, the logistic sigmoid activation function performed better than the hyperbola sigmoid. One explanation of this behavior is that ELM establishes different probability distributions with different activation functions. The different distributions of deep spatiotemporal features mapping employing different activation function are different, which directly affects the recognition rate of ELM. In other words, our experiment validates that logistic sigmoid is a more meaningful nonlinear mapping of deep spatiotemporal features than the hyperbola tangent.
- (ii) This parameter also depended upon the amount of data, as IAVID consists of 100 videos; therefore, a small number of hidden nodes are required. Whereas MuHAVI-Uncut and IXMAS dataset are larger than IAVID, the number of hidden nodes is greater.
- (iii) Adding more hidden nodes may not always result in better performance. However, the random selection of hidden nodes may lead the learning model to be suffering from overfitting or underfitting. Therefore, an incremental selection of hidden nodes will help obtain a better network model followed by pruning the unnecessary hidden nodes for better performance.

After the empirical selection of the type of decision function and number hidden nodes, these parameters will remain the same for the rest of the experiments.

**3.4. Quantitative Analysis of Proposed Technique.** IAVID dataset was evaluated using LOSO, LOAO and validation

TABLE 1: Performance of the proposed technique at different network depth on the IAVID-I dataset using LOAO validation scheme.

Deep spatiotemporal features	Features	No. Layers	Hidden Node	LOAO Accuracy %	Time Sec
$x_{17}$	1x4096	17	1300	67.98	0.00159
$x_{20}$	1x 4096	20	1300	46.25	0.00203
$x_{23}$	1x 1000	23	1300	43.10	0.00108
$x_{39}$	1x 4096	39	500	55.74	0.00059
$x_{42}$	1x 4096	42	500	44.96	0.001302
$x_{45}$	1x 1000	45	500	43.13	0.000811

TABLE 2: Performance of the proposed technique at the different network on the IAVID-I dataset using LOSO validation scheme.

Deep spatiotemporal features	Deep CNN layers	Features Dimension	Hidden Nodes	LOSO Accuracy %
$x_{17}$	17	1x 4096	500	79.75
			2100	82.19
			5000	81.62
			10000	81.37
$x_{39}$	39	1x 4096	500	64.93
			2100	77.76
			5000	76.43
			10000	74.96

training testing splits schemes and recorded recognition rates for 8 instructor's actions of 82.19%, 67.98%, and 81.43% using  $x_{17}$ .

It can be observed from Table 1 that deep spatiotemporal features  $x_{17}$  and  $x_{39}$  performed better as compared to other variants  $x_{20}$ ,  $x_{23}$ ,  $x_{39}$ ,  $x_{42}$ , and  $x_{45}$  computed from the same CNN model at LOAO scheme. This empirical analysis indicates that the shallower layers of network exhibit a deeper representation of action classes as compared to higher layers. At higher layers of CNN, some features are dropped out due to compression of representation using pooling and dropped-out layers. Moreover, the LOAO validation scheme illustrates the strength of the proposed technique against person independent HAR. As in LOAO all the action sequences of one actor is used as a testing and the remaining are used for training. This process is repeated for all the actors and the average performance of the system is recorded. Even when missing representation of one actor, the proposed technique has an accuracy of 67.98%.

From these results, we can infer that features  $x_{17}$  and  $x_{39}$  are reasonable choices to examine performance on LOSO validation scheme. The average recognition rate at LOSO is 82.19%.

From Table 2, it is observed that the number of hidden nodes is a significant parameter in tuning the performance of proposed technique, as similar deep spatiotemporal representations  $x_{17}$  and  $x_{39}$  generated different prediction rate at different numbers of hidden nodes, peaking at 2100. Increasing the hidden nodes further decreases accuracy, due to the different probability distribution of decision boundaries.

Similarly, from Table 3 it is observed that spatiotemporal deep representation  $x_{17}$  and  $x_{39}$  performed better than other

variants of deep spatiotemporal representation computed from the same CNN model. It can be observed from Table 3 that the recognition rate is higher when using  $x_{17}$  and  $x_{39}$  representations, for randomly sampled 70-30 training testing validation splits. The results of LOAO and LOSO on IAVID-I illustrate the strength of the proposed technique for activity recognition. We have compared the performance of the proposed technique with state-of-the-art methods and elaborated in detail in the comparison section.

From the confusion matrix for validation 70-30 split in Figure 5(a), the per class recognition rates for 'Interacting or idle', 'Pointing towards board or screen', 'Pointing students', 'Sitting', 'Using laptop', 'Using phone', 'Walking', and 'Writing on board' are 44.4%, 57.1%, 50%, 100%, 100%, 100%, 100%, and 100%, respectively. The average accuracy rate is 81.43%.

Similarly, from the confusion matrix for validation LOAO in Figure 5(b), the per class recognition rates for 'Interacting or idle', 'Pointing towards board or screen', 'Pointing students', 'Sitting', 'Using laptop', 'Using phone', 'Walking', and 'Writing on board' are 50%, 62.5%, 16.7%, 100%, 66.7%, 73.3%, 85.7%, and 88.9%, respectively. The instructor action class 'Pointing Board' achieved the lowest recognition rate because it is visually similar to action 'Interacting or idle', 'Pointing towards board or screen', 'Using phone'. The average accuracy of LOAO is 67.98%.

In case of LOSO, the per class recognition rate for 'Interacting or idle', 'Pointing towards board or screen', 'Pointing students', 'Sitting', 'Using laptop', 'Using phone', 'Walking', and 'Writing on board' is 91.7%, 80.0%, 85.7%, 61.5%, 100%, 100%, 78.6%, and 60%, as depicted in Figure 5(c). The average accuracy of LOSO is 82.19%.



TABLE 3: Performance of the proposed technique at different network depth on the IAVID-I dataset using 70-30 validation scheme.

Deep spatiotemporal features	Features	No. Layers	Hidden Node	Accuracy %	Time Sec
$x_{17}$	1x 4096	17	1300	81.43	0.00159
$x_{20}$	1x 4096	20	1300	78.34	0.00203
$x_{23}$	1x 1000	23	1300	73.98	0.00108
$x_{39}$	1x 4096	39	500	78.23	0.00059
$x_{42}$	1x 4096	42	500	75.88	0.001302
$x_{45}$	1x 1000	45	500	74.09	0.000811

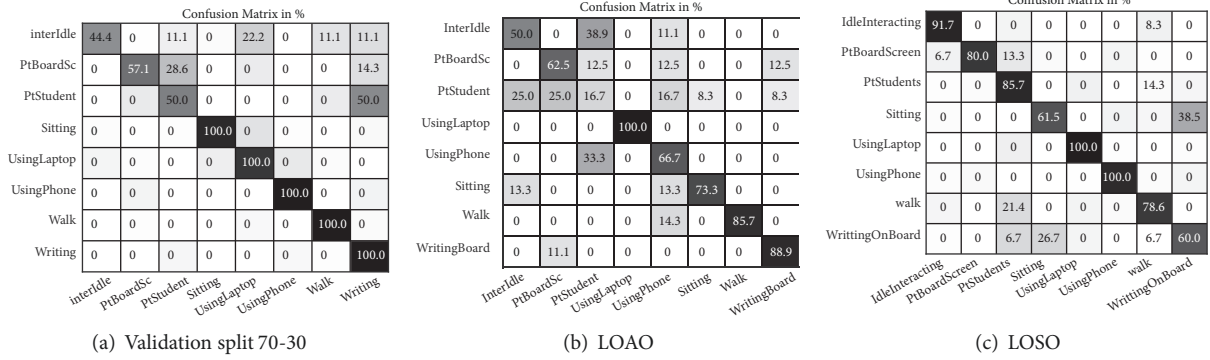
FIGURE 5: Confusion matrices of proposed technique computed from  $x_{17}$  on IAVID-I dataset.

TABLE 4: Performance comparison between backpropagation CNN model and feedforward proposed technique.

Deep spatiotemporal features	Validation Scheme	BP-CNN Accuracy %	Proposed technique Accuracy %
$x_{17}$	Splits	78.13	81.43
	LOAO	66.19	67.975
	LOSO	57.57	82.19
$x_{39}$	Splits	62.50	70.20
	LOAO	53.37	55.74
	LOSO	68.46	77.76

**3.4.1. Comparison between Backpropagation CNN Model and Feedforward Proposed Technique.** In this section, we elaborate on the effectiveness of our proposed technique as compared to CNN models. We have examined the performance of the proposed technique when deep spatiotemporal features  $x$  are used to train the backpropagation CNN model and feedforward ELM. The recorded results are presented in Table 3. The inference time for backpropagation CNN is higher than the proposed technique, which requires 1.5 milliseconds to recognize an activity; however, backpropagation CNN requires 50 msec (26.32 times slower). The backpropagation CNN model training and testing of 100 video sequences take 900 seconds; i.e., on average, it takes 0.05 seconds per sequence to recognize action at a frame rate of 1200 frames/second (FPS). However, our proposed technique recognizes actions at a frame rate of 40,000 FPS.

Backpropagation CNN also requires a considerable amount of time to reduce the training error during model learning depending on the amount of data. Our proposed technique performs better when we increase the number of ELM nodes, but the same is not true for backpropagation CNN. In backpropagation, the CNN model needed a large number of parameters tuning usually suffering the problem of overfitting when the amount of data is small like an IAVD-1 dataset. Therefore, we preferred to use smaller networks like Alexnet and VGG19 for computing spatiotemporal features from motion profiles. However, the proposed technique is able to extract deep features from any kind of CNN model. In our proposed scheme deep features are extracted from the CNN model without transfer learning in feedforward mode.

The results in Table 4 show that spatiotemporal features  $x_{17}$  and  $x_{39}$  recognize activities accurately when ELM is used as a classifier, compared to backpropagation CNN models. These results confirm that the proposed technique outperforms the backpropagation CNN models with respect to prediction accuracy and computational time.

**3.4.2. Comparison of Standard ELM with Variants of ELM.** We have evaluated the performance of the standard ELM classifier used in this paper against various variants of ELM classifier, i.e., minimum class variance ELM (MCV-ELM) [29], minimum variance ELM (MV-ELM) [30], self-adaptive evolutionary ELM (SADE-ELM) [28], and regularized ELM (R-ELM). The MVC-ELM [29] and MV-ELM [30] were introduced to improve the intraclass variance among the fine-grained activities and problems of unbalance data for activity recognition. In order to improve the intraclass

TABLE 5: Performance comparison of various variants of ELM using a deep spatiotemporal representation of action motion templates of IAVID-1 dataset.

Method	IAVID (70-30 split)		MuHAVi-Uncut (LOCO)	
	Accuracy	Computational Time	Accuracy	Computational Time
ELM	0.81250	6.05 m sec	82.04%	2280 sec
RELM [27]	0.81250	6.29 m sec	82.04%	2340 sec
SADE ELM [28]	0.235	0.9829 sec	50.98%	4080 sec
MCV-ELM [29]	0.1225	33.17sec	74.75%	5280 sec
MV-ELM [30]	0.1225	32.41 sec	74.75%	4380 sec

variation among the action classes, MCV-ELM and MV-ELM employed clustering based discriminant analysis and  $X^2$  distance within the action class scatter matrix. The SADE-ELM [28] is an evolutionary variant of ELM that optimized the hidden weight bias along with input weights through the differential evolutionary algorithm. While regularized ELM [27] explored the structural minimization of data outliers to reduce the problem of model overfitting without increasing the computational time.

To examine the behavior of standard ELM, MCV-ELM, MV-ELM, R-ELM, and SADE-ELM, the operational parameters remain the same for coherent evaluation. The deep spatiotemporal features  $x_{17}$  extracted from motion templates are used for model learning of standard ELM and other ELM variants. The number of hidden nodes  $L$  is chosen as 4096, as the feature dimension of  $x_{17}$  extracted from Alexnet's first fully connected layer is  $1 \times 4096$ . Therefore  $L$  is set as 4096.

The ELM and R-ELM cost parameter  $C_1$  and kernel parameter  $\gamma$  are determined empirically within  $[2e^{-7}, 2e^{-6}, \dots, 2e^{+7}]$  and  $[10e^{-7}, 10e^{-6}, \dots, 10e^{+7}]$ . The  $C_1$  is regulation coefficient presented to reduce the training error and norm of output weight, whereas  $\gamma$  is a constant that usually is greater than the norm of interconnection matrix of ELM hidden layer and bias vector [44]. Similarly, the operational cost parameter  $C_1$  and regression regularized parameter  $C_2$  for MCV-ELM and MV-ELM are empirically opted within  $[2e^{-7}, 2e^{-6}, \dots, 2e^{+6}, 2e^{+7}]$  and  $[10e^{-7}, 10e^{-6}, \dots, 10e^{+6}, 10e^{+7}]$ . The parameter  $C_2$  for MCV-ELM and MV-ELM is helpful in determining the output weights. These output weights are significant in estimating the trade-off between the training errors and training vector dispersion of the scatter matrix. Higher dispersion enables stronger decision boundaries and reduces outliers. The recognition performance of ELM and its variants is sensitive to the combination of regularization parameters  $C_1$ ,  $C_2$ , and  $\gamma$ . The optimal combination of parameters is dataset specific and achieved within the narrow range for model generalization.

The comparison presented in Table 5 highlights the effectiveness of the proposed approach. The performance of standard ELM for learning instructor action classes is comparable with R-ELM when high dimensional deep spatiotemporal action representation  $x_{17}$  is used for model learning. However, SADE-ELM, MCV-ELM, and MV-ELM performance is low for instructor action recognition using deep spatiotemporal action representation over IAVID-1 dataset. However,

the performance of MCV-ELM and MV-ELM is slightly low but comparable on MuHAVi-Uncut dataset using LOCO validation scheme. The reason for performance gain in case of MuHAVi-Uncut dataset as compared to the IAVID-1 dataset is that the number of samples comprising the MuHAVi-Uncut dataset classes is higher. However, it also highlights that these variants have higher dependency on the class sample rate as compared to the standard ELM. Moreover, the computational time (as presented in Table 5) for model learning of SADE-ELM, MCV-ELM, and MV-ELM is also higher than standard ELM and R-ELM, due to the generation of data-driven hidden node weights and bias. From Table 6 we can observe that the optimal combination  $(C_1, \gamma)$  for ELM and R-ELM is  $(2^{-6}, 10^{-6})$ , whereas the optimal combination  $(C_1, C_2)$  for MCV-ELM and MV-ELM is  $(2^3, 10^3)$  for better action recognition.

**3.4.3. Comparison with State-of-the-Art Methods.** In this section, we conclude the findings of our proposed technique as compared to the state-of-the-art techniques on the IAVID-1 dataset and also on publicly available multiview action recognition datasets (MuHAVi-Uncut and IXMAS), as illustrated in Table 7.

The proposed technique outperforms other techniques based on silhouettes in terms of precise recognition. To this end, we analyzed and compared our technique with methods [17, 20, 31, 32] on the IAVID-I dataset. The C3D features are computed from RGB IAVID-1 videos and produce 48.77% and 40% prediction accuracy using SVM and CNN. The performance of C3D features is comparable to 2D CNN without considering temporal information for HAR at frame level [18] Similarly, Bag of Expression [32] for HAR produces 26.67% recognition rate using handcrafted 3D-Harris and 3D-SIFT. Some recent silhouettes based HAR techniques are using MHIs described through HOG [20] and LBP-HOG [31] descriptor to recognize human activities through nearest neighbor and SVM classifiers. The instructor activity recognition rate for [20] is 63.5% and 55% for [31].

Since all the reported action recognition techniques based on MHI are using traditional features descriptor to represent the action, none of the reported techniques imply deep learning features to represent the spatiotemporal movement of the actor using MHIs. We believe that deep features are able to learn features in higher dimensions from motion templates that show better discriminative model learning for activity recognition. Table 7 shows comparative results

TABLE 6: Performance comparison of various variants of ELM at a different combination of the parameter on MuHAVi-Uncut dataset.

Dataset	Hidden nodes L	Regularization parameters			Accuracy			
		$\gamma$	Cost C1	Scatter Matrix C2	ELM	RELM	MCV-ELM	MVELM
MuHAVi-Uncut (LOCO validation scheme)	4096	$10^{-7}$	$2^{-7}$	$10^{-7}$	75.11%	75.11%	58.11%	NAN
		$10^{-6}$	$2^{-6}$	$10^{-6}$	82.04%	82.04%	59.76%	NAN
		$10^{-5}$	$2^{-5}$	$10^{-5}$	72.58%	72.58%	60.27%	24.49%
		$10^{-4}$	$2^{-4}$	$10^{-4}$	72.96%	72.96%	63.69%	65.79%
		$10^{-3}$	$2^{-3}$	$10^{-3}$	72.37%	72.37%	70.18%	73.84%
		$10^{-2}$	$2^{-2}$	$10^{-2}$	73.34%	73.34%	73.26%	73.52%
		$10^{-1}$	$2^{-1}$	$10^{-1}$	69.22%	69.22%	73.72%	73.84%
		$10^0$	$2^0$	$10^0$	72.55%	72.55%	73.61%	73.61%
		$10^1$	$2^1$	$10^1$	72.16%	72.16%	73.00%	73.00%
		$10^2$	$2^2$	$10^2$	71.16%	71.16%	73.74%	73.74%
		$10^3$	$2^3$	$10^3$	72.73%	72.73%	74.75%	74.75%
		$10^4$	$2^4$	$10^4$	73.28%	73.28%	73.39%	73.39%
		$10^5$	$2^5$	$10^5$	73.00%	73.00%	74.22%	74.23%
		$10^6$	$2^6$	$10^6$	71.72%	71.72%	73.33%	73.34%
		$10^7$	$2^7$	$10^7$	72.91%	72.91%	72.61%	72.61%

for the IAVID-I dataset using 70% training and 30% testing data.

Evidently, learned feature representation is beneficial for action recognition as the proposed technique outperforms the traditional feature representation, such that HOG [20, 31] confirms the benefits of good decision boundary among 8 instructor action classes using a feedforward network. These results confirm the benefits of deeply learned features for over traditional and deep learning action recognition techniques, as shown in Table 7.

Similar performance benefits are also obtained for the two standard multiview action recognition datasets, i.e., MuHAVI and IXMAS, as shown in Figure 6 and Table 7. We believe that the good performance of proposed technique on the MuHAVI video action recognition data is due to the flexibility of fusion of spatiotemporal motion profile with unsupervised deep learned features 'x' with feedforward network for learning model. The proposed technique improved the baseline recognition results on the MuHAVI-Uncut dataset in LOCO scheme by 29.84% and LOAO scheme by 9.56%. Similarly, in LOSO there is a slight improvement of 0.42%. However, on IXMAS dataset performance of the proposed technique is not so outstanding, due to the fact that actors of IXMAS do not have fixed angular positions towards the cameras. This characteristic introduces some misclassification in results. However, there is no significant variation in aspect within each view. The actions of IXMAS are visually fairly similar to each other like folding arms, watching watch and scratching head. The motion profile generated from these actions is therefore visually similar to each other. Therefore, the proposed technique does not precisely predict the action classes, though it does to some extent, as shown in Table 7.

To compare performance, we consider other approaches to MuHAVi-uncut and IXMAS dataset. For comparison we have implemented the approach [20, 33–35]. The operational parameters and noise removal technique across all the [20, 33–35] methods remain the same for fair comparison. In [20],  $\tau$  is the total number of video frames used for silhouettes generation, like our method. In [35], motion profiles were used to model the classifier for action recognition. The recognition rates of [35] using LOAO, LOCO and LOSO validation scheme are not more than 60%, while proposed approach significantly improves the recognition accuracy by 36.96%, 46.13%, 40.42% respectively as compared to [35]. In [33], action motion profiles were clustered through a self-organizing map (SOM) and clusters were further projected on manifold space to predict the action class through observable Markov Model. The technique proposed here performed 9.76% better than observable Markov Model at LOAO validation scheme, as shown in Table 7.

#### 4. Conclusion

In this paper, we presented a framework for instructor activity recognition by deep spatiotemporal features and feedforward Extreme Learning Machines by incorporating spatiotemporal instructor silhouettes information in single motion profile and representing them with high dimensional deep convolutional features. These deep spatiotemporal representations are used to learn the model for instructor activity recognition by employing an extreme learning machine as a classifier. The proposed scheme has shown several salient features including accurate prediction of instructor actions and performs recognition in feedforward fashion despite backpropagation

TABLE 7: Performance comparison of the proposed approach with state-of-the-art techniques.

Dataset	Validation scheme	Method	Accuracy
IAVID-1	Splits (70-30)	<i>Proposed technique</i>	81.43%
		C3D features with SVM classifier[17]	48.77%
		C3D features with CNN[17]	40.0%
		HOG representation of MHI with nearest neighbor classifier[20]	63.5%
		HOG and LBP representation of MHI with SVM classifier[31]	55%
		Harris 3D and HOG 3D with BOE[32]	26.67%
		Harris 3D, HOG/ HOF, BoF with MCV-ELM[29]	13.33%
		Harris 3D, HOG/ HOF, BoF with MV-ELM[30]	13.33%
MuHAVI-Uncut	LOAO	<i>Proposed technique</i>	93.66%
		HOG representation of MHI with nearest neighbor classifier[20]	84.1%
		Observable Markov model[33]	83.90%
		The sequence of key poses[34]	81.50%
		Learning discriminative key poses[35].	56.70%
	LOCO	<i>Proposed technique</i>	82.04%
		Deep spatiotemporal representation of MHI with MCV-ELM[29]	74.75%
		Deep spatiotemporal representation of MHI with MV-ELM[30]	74.75%
		HOG representation of MHI with nearest neighbor classifier[20]	52.2%
		The sequence of key poses [34]	50.4%
		Learning discriminative key poses [35].	31.4%
	LOSO	<i>Proposed technique</i>	97.02%
		HOG representation of MHI with nearest neighbor classifier[20]	96.6%
		The sequence of key poses [34]	86.5%
		Learning discriminative key poses [35].	56.6%
IXMAS	LOSO	<i>Proposed technique</i>	71.94%
		Substructure and boundary modeling [36]	76.5%
		Self-organizing map of action poses and fuzzy distance for MLP[37]	89.9%
		The sequence of key poses [34]	85.9%
		Multiview spatiotemporal histogram[38]	81.4%
	LOCO	Spatiotemporal volumes (3DSTVs) mapped to 4D[39]	78%
		<i>Proposed technique</i>	74.52%
		Spatiotemporal visual words to learn SVM model[40]	57.30%
		3D grid to learn HMM model for action recognition[41]	57.90%
		Sphere and rectangular feature trees with nearest neighbor classifier[42]	72.60%
		Histogram of silhouettes, horizontal and vertical optical-flow for action recognition[43]	58.10%

or iterative learning. Moreover, the proposed technique shows improvements in the challenges of scale, viewpoint variation, and multiple actors and accurately predicts actions. We have improved the baseline recognition rate on one of the multiview HAR datasets (MuHAVI-Uncut). In the future, we will explore new techniques to understand the classroom semantics for supporting instructor self-reflection mechanism for lecture effectiveness.

### Data Availability

We have used two publicly available datasets (MuHAVI and IXMAS). However, the IAVID-1 dataset used to support the findings of this study may be provided on request to

“Muhammad Haroon Yousaf”, who can be contacted at haroon.yousaf@uettaxila.edu.pk.

### Conflicts of Interest

The authors declare that they have no conflicts of interest.

### Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for research work carried in Centre for Computer Vision Research (C2VR) at University of Engineering and Technology Taxila, Pakistan. Sergio A Velastin acknowledges



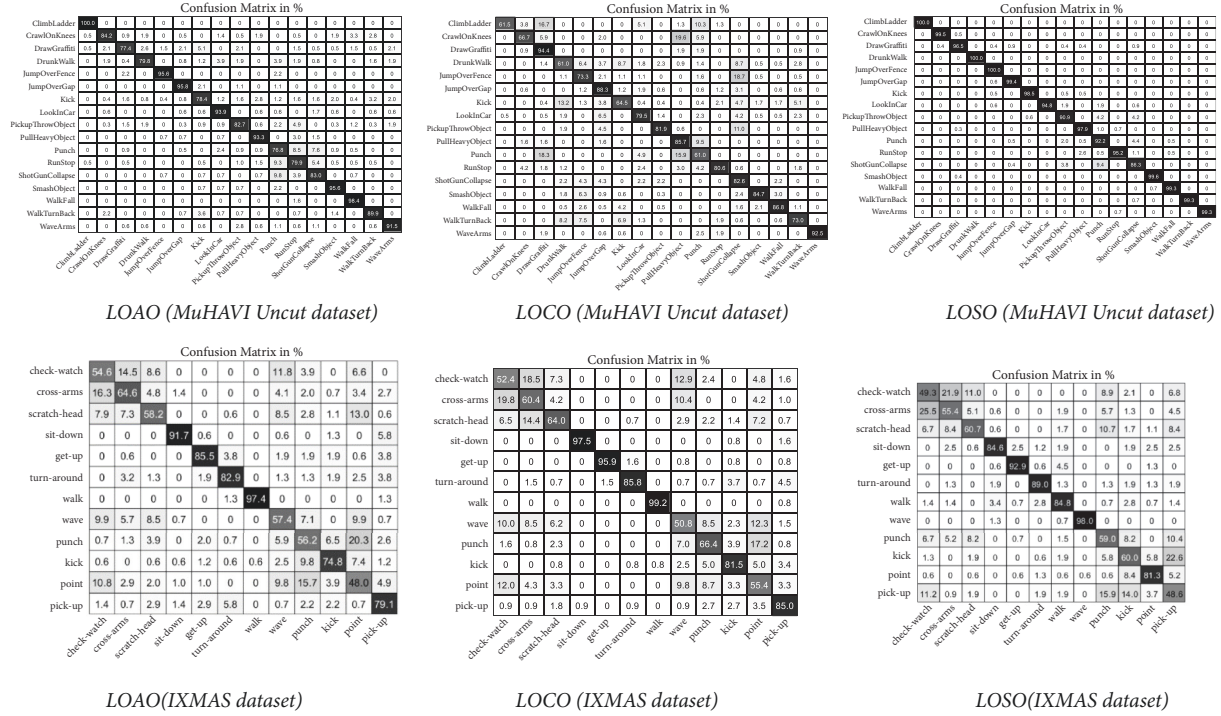


FIGURE 6: Confusion matrix of the proposed technique on MuHAVI-Uncut and IXMAS dataset.

funding by the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for Research, Technological Development and Demonstration under grant agreement no. 600371, el Ministerio de Economía y Competitividad (COFUND2013-51509), and Banco Santander. We are also very thankful to participants, faculty, and postgraduate students of Computer Engineering Department who took part in the data acquisition phase. Without their consent, this work was not possible.

## References

- [1] A. Raza, M. H. Yousaf, H. A. Sial, and G. Raja, "HMM-based scheme for smart instructor activity recognition in a lecture room environment," *The Smart Computing Review*, vol. 5, pp. 578–590, 2015.
- [2] M. H. Yousaf, K. Haroon, K. Ahmed, and H. A. Habib, "Human activity recognition for intelligent video lecture recording," in *Proceedings of the 14th International Conference in Recent Achievements in Mechatronics, Automation, Computer Science and Robotics*, pp. 101–109, Romania, 2010.
- [3] M. H. Yousaf, K. Azhar, and H. A. Sial, "A novel vision based approach for instructor's performance and behavior analysis," in *Proceedings of the Communications, Signal Processing, and their Applications (ICCSPA)*, pp. 1–6, 2015.
- [4] M. H. Yousaf, H. A. Habib, and K. Azhar, "Fuzzy classification of instructor's morphological features for autonomous lecture recording system," *Information-An International Interdisciplinary Journal*, vol. 16, pp. 6367–6382, 2013.
- [5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS '05)*, pp. 65–72, October 2005.
- [6] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.
- [7] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV '13)*, pp. 3551–3558, Sydney, Australia, December 2013.
- [8] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM International Conference on Multimedia (MM '07)*, pp. 357–360, September 2007.
- [9] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proceedings of the 19th British Machine Vision Conference (BMVC '08)*, vol. 275, pp. 1–10, September 2008.
- [10] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proceedings of the European conference on computer vision*, pp. 650–663, 2008.
- [11] J. Perš, V. Sulić, M. Kristan, M. Perše, K. Polanec, and S. Kovačič, "Histograms of optical flow for efficient representation of body motion," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1369–1376, 2010.
- [12] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.
- [13] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: a dataset of 101 human actions classes from videos in the wild," <https://arxiv.org/abs/1212.0402>, 2012.
- [14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition,"

- in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 2556–2563, November 2011.
- [15] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, “ActivityNet: a large-scale video benchmark for human activity understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, USA, June 2015.
  - [16] L. Wang, Y. Qiao, and X. Tang, “Action recognition with trajectory-pooled deep-convolutional descriptors,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 4305–4314, USA, June 2015.
  - [17] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.
  - [18] F. Zhu, L. Shao, J. Xie, and Y. Fang, “From handcrafted to learned representations for human action recognition: a survey,” *Image and Vision Computing*, vol. 55, pp. 42–52, 2016.
  - [19] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014, NIPS 2014*, pp. 568–576, Canada, December 2014.
  - [20] F. Murtaza, M. H. Yousaf, and S. A. Velastin, “Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description,” *IET Computer Vision*, vol. 10, no. 7, pp. 758–767, 2016.
  - [21] M. G. Sherin and E. B. Dyer, “Teacher self-captured video: learning to see,” *Phi Delta Kappan*, vol. 98, no. 7, pp. 49–54, 2017.
  - [22] S. A. Nagro, L. U. deBettencourt, M. S. Rosenberg, D. T. Carran, and M. P. Weiss, “The effects of guided video analysis on teacher candidates’ reflective ability and instructional skills,” *Teacher Education and Special Education*, vol. 40, no. 1, pp. 7–25, 2017.
  - [23] N. R. Howe and A. Deschamps, “Better foreground segmentation through graph cuts,” arXiv preprint cs/0401017, 2004.
  - [24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12)*, pp. 1097–1105, Lake Tahoe, Nev, USA, December 2012.
  - [25] N. Liu and H. Wang, “Ensemble based extreme learning machine,” *IEEE Signal Processing Letters*, vol. 17, no. 8, pp. 754–757, 2010.
  - [26] G. B. Huang, Q. Y. Zhu, and C. K. Siew, “Extreme learning machine: a new learning scheme of feedforward neural networks,” in *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 2, pp. 985–990, July 2004.
  - [27] W. Y. Deng, Q. H. Zheng, and L. Chen, “Regularized extreme learning machine,” in *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM '09)*, pp. 389–395, April 2009.
  - [28] J. Cao, Z. Lin, and G.-B. Huang, “Self-adaptive evolutionary extreme learning machine,” *Neural Processing Letters*, vol. 36, no. 3, pp. 285–305, 2012.
  - [29] A. Iosifidis, A. Tefas, and I. Pitas, “Minimum class variance extreme learning machine for human action recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1968–1979, 2013.
  - [30] A. Iosifidis, A. Tefas, and I. Pitas, “Minimum variance extreme learning machine for human action recognition,” in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2014*, pp. 5427–5431, Italy, May 2014.
  - [31] M. Ahad, M. Islam, and I. Jahan, “Action recognition based on binary patterns of action-history and histogram of oriented gradient,” *Journal on Multimodal User Interfaces*, vol. 10, no. 4, pp. 335–344, 2016.
  - [32] S. Nazir, M. H. Yousaf, J.-C. Nebel, and S. A. Velastin, “A Bag of Expression framework for improved human action recognition,” *Pattern Recognition Letters*, vol. 103, pp. 39–45, 2018.
  - [33] C. Orrite, M. Rodriguez, E. Herrero, G. Rogez, and S. A. Velastin, “Automatic segmentation and recognition of human actions in monocular sequences,” in *Proceedings of the 22nd International Conference on Pattern Recognition, ICPR 2014*, pp. 4218–4223, Sweden, August 2014.
  - [34] A. A. Chaaoui, P. Climent-Pérez, and F. Flórez-Revuelta, “Silhouette-based human action recognition using sequences of key poses,” *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.
  - [35] S. Cheema, A. Eweiri, C. Thureau, and C. Bauckhage, “Action recognition by learning discriminative key poses,” in *Proceedings of the Proceeding of the IEEE International Conference on Computer Vision Workshops (ICCV '11)*, pp. 1302–1309, Barcelona, Spain, November 2011.
  - [36] Z. Wang, J. Wang, J. Xiao, K.-H. Lin, and T. Huang, “Substructure and boundary modeling for continuous action recognition,” in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*, pp. 1330–1337, USA, June 2012.
  - [37] A. Iosifidis, A. Tefas, and I. Pitas, “View-invariant action recognition based on artificial neural networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 412–424, 2012.
  - [38] G. Srivastava, H. Iwaki, J. Park, and A. C. Kak, “Distributed and lightweight multi-camera human activity classification,” in *Proceedings of the 2009 Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, pp. 78–81, Como, Italy, August 2009.
  - [39] P. Yan, S. M. Khan, and M. Shah, “Learning 4D action feature models for arbitrary view action recognition,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008*, pp. 1–7, 2008.
  - [40] C. Huang, Y. Yeh, and Y. F. Wang, “Recognizing actions across cameras by exploring the correlated subspace,” in *Computer Vision – ECCV 2012. Workshops and Demonstrations, Lecture Notes in Computer Science*, pp. 342–351, Springer, Berlin, Germany, 2012.
  - [41] D. Weinland, E. Boyer, and R. Ronfard, “Action recognition from arbitrary views using 3D exemplars,” in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–7, October 2007.
  - [42] K. K. Reddy, J. Liu, and M. Shah, “Incremental action recognition using feature-tree,” in *Proceedings of the 12th International Conference on Computer Vision, ICCV 2009*, pp. 1010–1017, Japan, October 2009.
  - [43] A. Farhadi and M. K. Tabrizi, “Learning to recognize activities from the wrong view point,” in *Computer Vision*, vol. 5302 of *Lecture Notes in Computer Science*, pp. 154–166, Springer, Berlin, Germany, 2008.
  - [44] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, “Extreme learning machine for regression and multiclass classification,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

